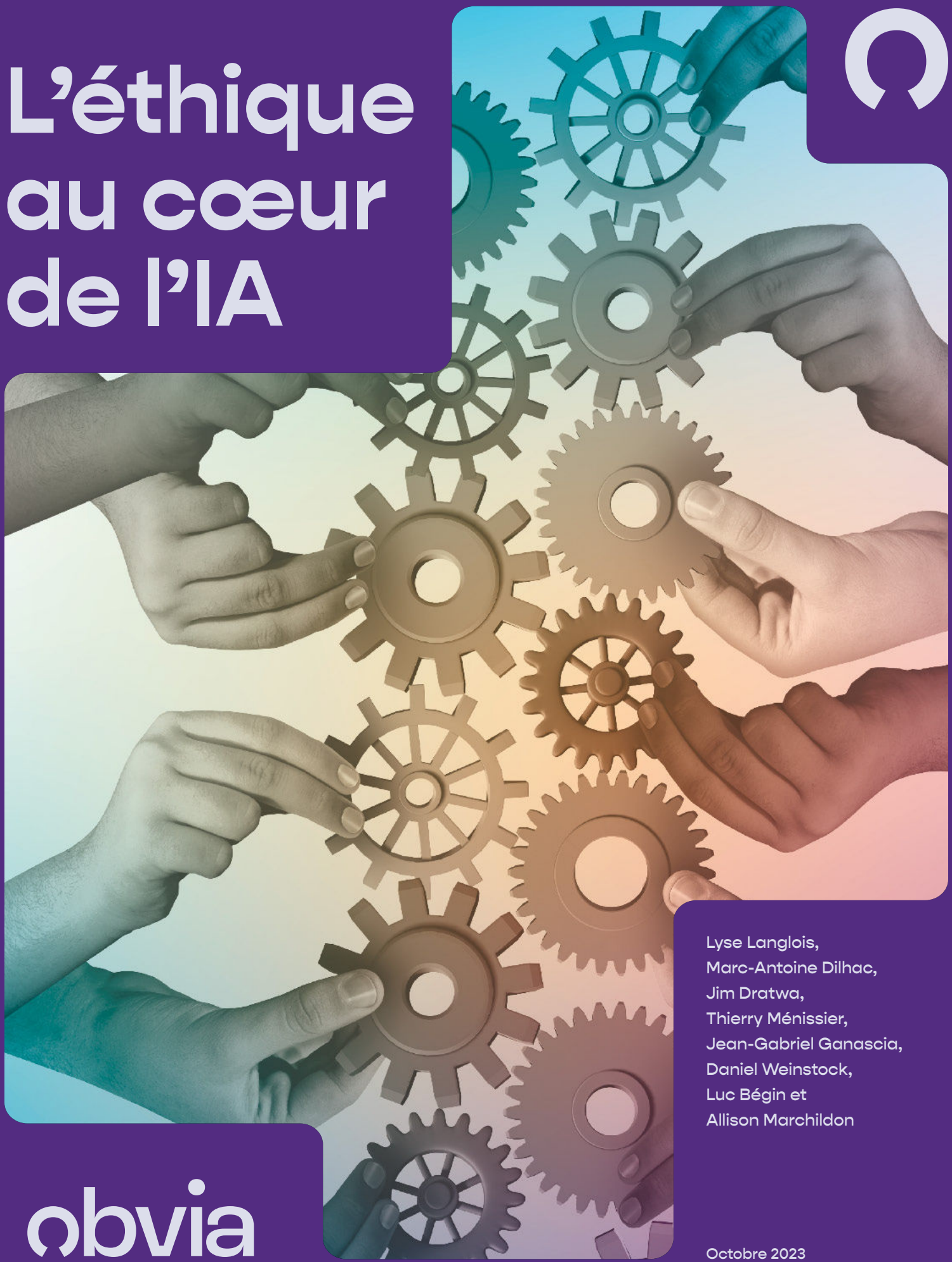
# L'éthique au cœur de l'IA



Lyse Langlois,
Marc-Antoine Dilhac,
Jim Dratwa,
Thierry Ménissier,
Jean-Gabriel Ganascia,
Daniel Weinstock,
Luc Bégin et
Allison Marchildon

obvia

Octobre 2023

# Note introductive

Issu d'un travail collaboratif regroupant des spécialistes de l'éthique, de la philosophie, de l'informatique et de l'économie, le présent document vise à préciser et clarifier le rôle que doit occuper l'éthique à l'ère de l'intelligence artificielle (IA), et à mettre en lumière comment cette notion peut être appliquée et mise en œuvre de manière efficace et fructueuse. S'adressant à l'ensemble des individus engagés, de près ou de loin, dans le développement de l'IA, ce document met de l'avant une éthique centrée sur la réflexivité et le dialogue. Dans une volonté de traduire plus concrètement cette vision, il met en lumière l'approche méthodologique utilisée pour construire la Déclaration de Montréal et propose également quelques pistes de recommandation. En somme, le présent texte plaide pour l'inclusion d'une réelle réflexion éthique dans l'ensemble des étapes du processus de développement de l'IA. Il se veut ainsi une main tendue, un appel à la collaboration entre éthiciennes et éthiciens, développeuses et développeurs et membres de l'industrie afin de véritablement **intégrer l'éthique au cœur de l'IA**.


## Autrices et auteurs :

**Lyse Langlois**, Ph.D, professeure au département de relations industrielles de l'Université Laval et directrice générale de l'OBVIA

**Marc-Antoine Dilhac**, Ph.D, professeur au département de philosophie de l'Université de Montréal et directeur scientifique-gouvernance et collaboration internationale de l'OBVIA

**Jim Dratwa**, Ph.D, conseiller à la Commission européenne et directeur de l'équipe en éthique des sciences et des nouvelles technologies

**Thierry Ménissier**, Ph.D, professeur de philosophie à l'Université Grenoble Alpes, directeur de la chaire éthique et IA de l'institut grenoblois MIAI (PIA 3IA)

**Jean-Gabriel Ganascia**, Ph.D, professeur d'informatique à la faculté des sciences de Sorbonne Université, président du comité d'éthique du CNRS

**Daniel Weinstock**, Ph.D, professeur de philosophie au département de droit de l'Université McGill

**Luc Bégin**, Ph.D, professeur à la faculté de philosophie de l'Université Laval

**Allison Marchildon**, Ph.D, professeure au département de philosophie et d'éthique appliquée de l'Université de Sherbrooke et co-responsable de l'axe éthique, gouvernance et démocratie de l'OBVIA

## Soutien à la coordination et à la recherche :

**Félix-Arnaud Morin-Bertrand**, M.A, professionnel de recherche à l'OBVIA



## Remerciements :

**Nathalie de Marcellis-Warin**, Ph.D, professeure au département de mathématiques et de génie industriel à Polytechnique Montréal, Présidente-directrice générale du CIRANO, directrice scientifique-Outils de mesure, veille et enquêtes de l'OBVIA

Produit avec le soutien financier des Fonds de recherche du Québec

Québec

Fonds de recherche – Nature et technologies
Fonds de recherche – Santé
Fonds de recherche – Société et culture

ISBN: 978-2-925138-26-6

# Table des matières

# Introduction

L'intelligence artificielle (IA) occupe aujourd'hui une place sans précédent dans la vie des êtres humains. En effet, les algorithmes d'IA sont désormais intégrés à de multiples sphères de nos vies et de nos sociétés, et influencent de ce fait nos choix, nos relations avec les autres ainsi que notre manière de travailler et d'apprendre. Plus largement, les technologies numériques qui fondent l'IA, modifient l'organisation de notre monde et, simultanément, notre manière d'appréhender et de comprendre ce dernier (Floridi, 2023). L'éthique a été désignée et employée au cours des dernières années pour réfléchir aux inquiétudes et enjeux liés au déploiement croissant de l'IA dans la société (Koniakou, 2023). L'intérêt grandissant envers l'éthique a permis de dégager des principes éthiques en IA largement reconnus et même de définir une norme internationale (UNESCO, 2021). Malgré cette étape importante, **beaucoup reste à faire pour démystifier l'éthique et les retombées qu'elle peut générer sur le plan des compétences individuelles, organisationnelles et stratégiques.** Le déploiement des modèles d'**intelligence artificielle générative** suscite actuellement des inquiétudes et soulève de nouveaux enjeux en matière d'éducation, de droit, de travail, de culture et de démocratie. Dans ce contexte, l'éthique devient incontournable, car sa vocation est éminemment pratique. Sa place, sa fonction et son exercice méritent d'être vus comme une composante essentielle du déploiement d'une capabilité morale (Nussbaum, 2012) nécessaire pour mener une vie humaine digne et libre dont la finalité est le progrès humain.

L'éthique est un concept humain complexe enraciné dans les normes culturelles, les valeurs et les croyances sociales. Son intégration dans les processus algorithmiques doit reposer sur un équilibre entre différentes valeurs et la compréhension des contextes. C'est grâce à cet équilibre que peut être menée une réflexion articulée sur le plan du raisonnement moral. Les **systèmes d'intelligence artificielle (SIA)** prennent des décisions basées sur des algorithmes et des données, qui peuvent toutefois être imparfaits ou biaisés. L'éthique fait intervenir le pouvoir de faire, le pouvoir de dire, le pouvoir de raconter et l'imputabilité (Ricœur, 2004). Ces capacités valorisent la responsabilité de chacun à l'égard de lui-même, des autres et du monde. Cette dimension de l'éthique est très importante, et ce, particulièrement dans la révolution technologique qu'est l'IA. Il nous apparaît donc nécessaire de la clarifier afin à la fois d'en mesurer toute la complexité et de révéler l'apport que l'éthique peut représenter en situation de choix.

« L'éthique est un concept humain complexe enraciné dans les normes culturelles, les valeurs et les croyances sociales. Son intégration dans les processus algorithmiques doit reposer sur un équilibre entre différentes valeurs et la compréhension des contextes. »

# Contexte

L'année 2022 a marqué un tournant dans le monde de l'IA. Le perfectionnement et le déploiement de grands modèles de langage tels que ChatGPT ont eu l'effet d'une véritable onde de choc dans le domaine de l'IA ainsi que dans l'ensemble de la société. Dès leur arrivée, l'ampleur et l'étendue des capacités génératives de ces modèles ont été remarquées, et plusieurs personnes en sont venues à envisager les conséquences potentielles de leur déploiement à grande échelle. Ce sont en effet les capacités génératives des grands modèles de langage comme GPT (OpenAI), Claude (Anthropic), PaLM (Google) ou Llama (Meta) qui retiennent l'attention et expliquent l'utilisation fréquente du terme *IA générative*. Ces grands modèles de langage, désormais multimodaux, sont en mesure d'effectuer, presque instantanément, des tâches complexes et variées allant de la rédaction de textes à la génération d'images en passant par l'écriture de codes informatiques. ChatGPT n'était à cet égard que la pointe de l'iceberg que représente l'IA générative (Peres et al., 2023). La mise à l'épreuve de ces modèles avec certaines évaluations professionnelles a révélé les possibilités, mais aussi les limites des modèles d'évaluation professionnels actuels. Cette percée technologique modifie la façon dont nous interagissons avec les ordinateurs, mais demande aussi que nous revoyions nos façons de faire, ce qui inaugure une nouvelle ère d'interaction humain-machine. Le perfectionnement des grands modèles de langage a aussi donné lieu à d'innombrables débats, réflexions et prises de position au sein des milieux universitaires et de l'industrie concernant l'avenir des développements de l'IA. Les nouveaux modèles d'IA générative simulent de plus en plus certaines parties du fonctionnement du cerveau et pour certaines capacités, ils offrent des performances supérieures à ce qu'un humain peut faire. Des auteurs affirment même qu'ils pourraient atteindre une forme de « moralité artificielle » (Butlin et al., 2023).

« Cette percée technologique modifie la façon dont nous interagissons avec les ordinateurs, mais demande aussi que nous revoyions nos façons de faire, ce qui inaugure une nouvelle ère d'interaction humain-machine. »

**Toutes et tous reconnaissent qu'une étape importante a été franchie entraînant autant d'inquiétudes que de promesses d'amélioration.** Plusieurs acteurs de l'industrie et même des milieux universitaires ont par ailleurs conjointement pris position en faveur d'une prise en charge accrue des risques de l'IA, considérés comme des risques d'extinction au même titre que les pandémies et la guerre nucléaire (Center for AI Safety, 2023). Ce genre de position peut sans aucun doute être qualifié de catastrophiste, et il ravive certainement le débat entre vision inflationniste et déflationniste de l'IA (Maclure, 2020). Récemment, la vision inflationniste de l'IA a grandement été portée par différents acteurs participant au développement de cette technologie. Certaines motivations derrière cette prise de position ne sont pas sans rappeler l'influence des mouvements tels que le **long-termisme**, le **transhumanisme,** ou **le singularitarisme** technologiques. Ces courants accumulent les paradoxes, se contredisent et se réclament pour la plupart de la science, mais ils relèvent le plus souvent d'un mythe (Cassiani-Laurin, 2018). Derrière ces approches souvent entretenues et promues par l'écosystème technologique se retrouve une symbolique associée à des scénarios catastrophiques ou à une forme de religiosité garantissant la promesse d'un salut virtuel infini. Ces courants influencent fortement le paradigme technologique actuel en se liant aux idéaux scientifiques, techniques et sociaux de notre époque.

## La complexité du social et la production d'éthiques

La place que prennent les développements technologiques dans nos vies bouleverse les activités humaines et influence notre condition et notre environnement social, ce qui entraîne du même coup une complexité sociale. Il est parfois difficile d'en voir clairement tous les effets sociaux, car l'IA façonne notre présent et notre avenir plus rapidement que toute autre invention. Mener une réflexion approfondie sur ces effets en suivant le même tempo que celui du développement de l'IA s'avère donc complexe. Les changements dépendent des modes d'appropriation des populations et demandent qu'une vigilance permanente soit exercée concernant la manière dont les usages des technologies évoluent, s'implantent et s'articulent au sein même de ces populations. Il est déjà possible de percevoir que dans différentes cultures sociétales, **une nouvelle vision du monde et de nouveaux rapports avec les humains et la biosphère sont en train de s'installer.** Ces relations sont d'une nature complètement différente des rapports sociaux que nous avions auparavant. Un peu comme la naissance de la bioéthique a fait suite aux avancées en matière de biologie et de médecine, nous assistons en ce moment à l'apparition d'une éthique spécifique appliquée à l'IA.

Comment cerner l'éthique, en quoi devrait-elle consister et comment pourrait-elle être propre au domaine de l'IA ? Malgré sa grande popularité dans ce domaine depuis 2016, dans les faits, l'éthique a été malmenée, réduite, parfois rendue impuissante à exercer un rôle de premier plan dans le développement de l'IA. Tantôt dévoyée, dénaturée, instrumentalisée, l'éthique a fait l'objet de nombreux malentendus et a été souvent détournée à des fins stratégiques et marketing au profit d'une idéologie productiviste. Le rapport du *European Group on Ethics in Science and New technologies* intitulé *Artificial Intelligence, Robotics and 'Autonomous' Systems* (EGE, 2018 : 14) rappelle qu'en l'absence d'une réflexion coordonnée sur les enjeux éthiques et sociaux de l'IA, le risque de verser dans un marché de principes et valeurs pour justifier en surface les conduites attendues est réel. Ceci n'est pas sans rappeler les deux premières vagues de tentatives de formalisation de l'éthique de l'IA. Ces étapes ont pourtant été utiles à la définition de principes pouvant encadrer les systèmes d'IA (Georgieva & al., 2022) puisqu'elles ont fait émerger certains consensus sur les principes à prioriser. Toutefois, certaines et certains ont affirmé que cet intérêt envers l'éthique demeurait factice et qu'il était plutôt révélateur de stratégies d'industries visant à contrer toute réglementation. D'autres ont mis en évidence un certain relativisme interprétatif qualifié d'*ethics shopping* et ont ainsi démontré une sorte d'incompatibilité des normes réduisant les chances de comparaison, de concurrence et de responsabilité (Floridi & Clément-Jones, 2019). Ces premières vagues ont contribué à la prédominance actuelle d'une éthique de l'IA qu'on pourrait qualifier de minimaliste (Ménissier, 2023) et dont l'ambition s'est surtout limitée à la prévention et à la limitation des risques. Devant ce constat, il convient de recadrer cette production d'éthiques diverses, de dégager le sens de cette production en contexte de développement technologique et les motifs qui président à sa constitution, dans l'objectif de placer l'éthique au cœur de l'IA.

« Tantôt dévoyée, dénaturée, instrumentalisée, l'éthique a fait l'objet de nombreux malentendus et a été souvent détournée à des fins stratégiques et marketing au profit d'une idéologie productiviste. »

# 1
# L’éthique et l’éthique de l’IA

*Dans quel monde voulons-nous vivre ? ; À quelles valeurs sommes-nous attachés en tant que communauté humaine ? ; Comment doit-on construire ce « nous » ?*
(Dratwa, 2019)

# 1. L'éthique et l'éthique de l'IA

La popularité de l'éthique dans différents milieux professionnels et dans le public réside dans la perception qu'elle est un discours moral aisément accessible, qu'on peut tenir souvent de manière intuitive et sans devoir d'abord faire une réflexion approfondie. L'éthique est ainsi reconnue à la fois comme une discipline philosophique et comme une activité qui s'adresse à tout le monde sans prérequis, sans formation philosophique, par exemple. Ce paradoxe n'est qu'apparent et il est aisé de le dissiper. *L'éthique, c'est faire le bien et éviter le mal* : ce type d'affirmation populaire est courant, et comme tout le monde aurait une idée de ce qu'est le bien, tout le monde pourrait faire de l'éthique tout simplement en respirant. Pourtant, distinguer le bien du mal n'est pas du tout évident et l'éthique ne peut être réduite qu'à cela. Il est vrai que l'éthique contient l'idée d'un comportement adéquat, comportement qui peut être bon ou conforme à une valeur ou à une règle (un principe). Pour connaître ou déterminer les valeurs ou les règles du bon comportement, on fait appel aux notions de finalité morale, de bien, mais aussi de devoir, d'obligation, de décision, de dilemme, de vertu, etc. Or ces notions ne sont pas intuitives.

## 1.1. L'éthique contre les listes à cocher

Depuis 2017, en particulier avec la création des « 23 principes d'Asilomar sur l'IA » et la Déclaration de Montréal pour un développement responsable de l'IA, l'éthique de l'IA a pris la forme de listes de principes et de règles. Nous reviendrons, dans la section 3.2, sur la Déclaration de Montréal, dont l'ambition était moins de donner une « recette » pour faire de l'IA éthique que d'inviter à une pratique réflexive et délibérative. Cependant, cette déclaration et toutes celles qui l'ont suivie ont été présentées sous la forme fixe, statique, du catalogue d'énoncés vertueux. Ainsi, pour faire de l'éthique de l'IA, il suffirait d'établir la bonne liste de principes (3, 5, 7, 9 ou 10 principes) et de la suivre à la lettre. Il ne resterait qu'à traduire ensuite la liste de principes en liste d'évaluation et de contrôle. La forme du catalogue des 10 commandements ou de la liste à cocher est séduisante parce qu'elle est simple et qu'elle donne le sentiment d'accéder à tout ce qu'il faut savoir pour agir bien sans devoir entamer une réflexion profonde. Néanmoins, **l'éthique ne peut pas tenir dans une liste.**

« ...l'éthique est une réflexion argumentée qui est soumise à des procédures de raisonnement rationnel. »

L'éthique, c'est d'abord une *réflexion morale raisonnée*, donc qui s'appuie sur des raisons. Cette façon de réfléchir s'est imposée dès l'Antiquité comme une des branches fondamentales de la philosophie. L'éthique a pour tâche de déterminer les règles de vie et d'action, de donner des recommandations, voire des injonctions pour bien vivre. La plupart des gens ont une certaine habitude de cette manière de faire de l'éthique : c'est celle qu'on retrouve dans des formules populaires comme *carpe diem* (littéralement, « cueille le jour » ou « profite du temps présent »). D'autres types de règles éthiques existent également, comme les devoirs moraux : « on ne doit pas mentir », « ne fais pas à autrui ce que tu ne voudrais pas qu'on te fasse », etc. Ces règles aussi sont très communes, trouvant leur source autant dans les textes religieux que dans la culture populaire. Cependant, l'éthique, ce n'est pas ces règles (ou la liste de ces règles). Ce qui distingue l'éthique comme discipline philosophique, d'un domaine disciplinaire comme le droit, réside dans sa démarche : l'éthique est une réflexion argumentée qui est soumise à des procédures de raisonnement rationnel. C'est quand on s'interroge sur les fondements rationnels d'une règle morale (qu'elle soit religieuse ou non), sur ses conséquences logiques, sur sa place dans un système de règles cohérent qu'on entre dans le domaine de l'éthique comme discipline philosophique. De ce point de vue, établir des listes de principes, de normes, est une activité annexe, voire mineure, de l'éthique.

## ÉTHIQUE ET DROIT

L'éthique et le droit sont parfois confondus, parfois mis en opposition. Ils ont des univers normatifs et dispositifs distincts, mais partagent un langage commun, soit les règles et les normes de la conduite humaine. À la différence du droit, l'éthique est incitative et se réfère à l'identification et à l'expression de valeurs ou principes devant orienter une action. Les valeurs et principes, lorsqu'ils sont partagés, permettent d'orienter l'action et d'apporter la justification sociale nécessaire dans ce travail d'analyse. Sa force réside dans l'interprétation et engage une forme d'appropriation des acteurs dans la construction du sens de la norme (Verhaegen, 1984). Elle engage aussi la responsabilité des acteurs qui sont interpellés par ce travail d'appropriation ancrée dans les situations concrètes. Quant au droit, il s'inscrit dans la voie de l'obligation et des devoirs enchâssés dans les lois et règlements. C'est le pouvoir des règles comme mode de détermination de l'action humaine. La finalité du droit est la conformité de la conduite aux règles juridiques, le tout approuvé par un seul acteur externe, à savoir un acteur public (juge ou législateur) qui sanctionne la mauvaise conduite. Le registre des deux notions est différent, ne doit pas être confondu ni réduit par des termes souvent popularisés comme le *droit mou* versus le *droit dur* les vidant de leur substance première.

### 1.2. Les normes éthiques et les normes techniques

En philosophie, le concept de « norme » désigne un énoncé qui indique ce que l'on doit faire ; il peut servir comme synonyme de *règle*. Dans le domaine industriel et dans celui de l'IA en particulier, le terme de « norme » est souvent utilisé pour désigner des normes techniques. Norme éthique ou norme technique, les deux types de normes renvoient donc à ce que l'on doit faire. Mais selon Aristote, une différence cruciale sépare la norme de l'éthique de celle de la technique : leur relation à ce qui est produit.

« Concevoir des systèmes d'IA fiables, ce n'est pas faire de l'éthique, c'est simplement répondre à une exigence technique… »

Dans le cas des règles éthiques, l'action ne produit rien d'extérieur à l'agent (l'entité qui agit). La valeur de l'action ne réside pas dans un produit, mais dans l'action elle-même : l'action généreuse est bonne en elle-même, même si elle a pour effet d'aider une personne malintentionnée. Au contraire, dans le cas des normes techniques, la valeur de l'action réside dans le produit. Ainsi, le fait de concevoir un système d'IA robuste, stable, « sécuritaire » n'a pas de valeur en soi : c'est la réalité (le produit) du système d'IA robuste, stable ou « sécuritaire » qui a de la valeur. Concevoir des systèmes d'IA fiables, ce n'est pas faire de l'éthique, c'est simplement répondre à une exigence technique qui est la base du travail des ingénieures et ingénieurs. Dans un même ordre d'idées, la cybersécurité n'est pas un problème éthique, mais technique.

Il est vrai qu'une informaticienne ou un informaticien qui construirait un système d'IA au mépris des normes techniques de sécurité agirait mal, en contradiction avec des engagements éthiques qui seraient ceux du respect des procédures, de la vigilance, de l'honnêteté, etc. Toutefois, la norme de sécurité ou de fiabilité est technique, elle n'est pas en elle-même éthique. Faire de l'éthique impose une réflexion sur ce qui a de la valeur morale, sur ce qui donne du sens à nos actions, à notre vie commune, sur les finalités désirables ou justes, mais aussi sur ce qui nous définit comme des êtres (des agents) moraux. En réfléchissant aux règles adéquates de l'action et du comportement, les philosophes se sont efforcés de les justifier rationnellement et ils ont élaboré des systèmes complexes de principes et de valeurs. C'est ce qu'on appelle les « doctrines morales ».

## LES DOCTRINES MORALES POUR PENSER LES *MORAL MACHINES*

Traditionnellement, les doctrines éthiques (ou morales) sont classées en trois grandes familles : l'éthique de la vertu, l'éthique conséquentialiste et l'éthique déontologique. L'éthique de la vertu (ou éthique arétique[1]) attribue la valeur de l'action au caractère de l'agent : l'action généreuse d'une personne a de la valeur parce qu'elle témoigne du caractère généreux, de la vertu de cette personne. L'éthique conséquentialiste détermine la valeur de l'action en évaluant ses conséquences : plus les conséquences d'une action sont bonnes selon une finalité choisie, plus l'action est bonne. Par exemple, selon une pensée utilitariste, une action est bonne si elle maximise le plaisir ou le bien-être. Enfin, l'éthique déontologique rapporte la valeur de l'action au respect des règles ou de l'autonomie des personnes, même si cela engendre de mauvaises conséquences. Par exemple, l'action de ne pas mentir ne doit pas être motivée par la volonté de maximiser les bonnes conséquences (et d'éviter les mauvaises), mais uniquement par le respect du devoir de dire la vérité ou par le respect pour la personne.

Cette manière de présenter l'éthique est très populaire en philosophie et au-delà de la philosophie. En éthique de l'IA, cette triade éthique sert parfois dans des cas pratiques, notamment dans la programmation des véhicules autonomes éthiques : les véhicules doivent-ils « choisir » l'action qui augmentera le bien-être (ou minimisera les torts) des personnes (éthique conséquentialiste) ou bien celle qui respectera leur dignité (éthique déontologique) ? Cette question est ainsi soulevée quand un agent se trouve face à un choix tragique, un dilemme où quoi qu'il choisisse, la vie de personnes sera affectée négativement. Par exemple, dans la situation où un accident est inévitable et où le véhicule peut poursuivre sa route et tuer X personnes ou virer sur une autre voie et tuer Y personnes, la question sera alors de savoir si X et Y, deux nombres, sont le seul facteur à prendre en compte ou si le fait d'impliquer des personnes dans l'accident alors qu'elles étaient sur une autre voie est acceptable moralement. On reconnaît là le fameux dilemme du tramway (the trolley problem) (Foot, 1967/2002), qui a été popularisé par le projet Moral Machine du MIT, où l'agent moral n'est pas une personne, mais un véhicule autonome (Awad, Dsouza, Kim, et al., 2018).

1 La vertu se dit *arêtê* en grec, d'où l'adjectif arétique pour désigner l'éthique de la vertu

# 2
# L’éthique de l’IA

*Ainsi, face aux interrogations légitimes des informaticiennes et informaticiens sur la manière d’appliquer les principes, une seule réponse est possible : il faut réfléchir en situation et se garder de vouloir automatiser le raisonnement moral.*

# 2. L'éthique de l'IA

Les paragraphes précédents ont permis de clarifier la notion d'éthique et de mettre en évidence plusieurs applications possibles de l'éthique au domaine de l'IA. Ainsi, la distinction entre normes éthiques et normes techniques est fondamentale pour cerner le champ d'application de l'éthique en IA. De plus, l'esquisse des trois familles de doctrines éthiques a montré qu'un problème éthique en IA peut être abordé de plusieurs manières. Mais est-ce que l'éthique de l'IA est tout simplement de l'éthique traditionnelle appliquée à l'IA ?

## 2.1. L'éthique appliquée à l'IA

En philosophie, l'éthique (1) et l'éthique appliquée (2) sont distinguées d'une manière qui semble très intuitive à première vue : l'éthique appliquée (2) est l'éthique (1) appliquée à une pratique sociale (par exemple, la guerre), à un secteur d'activité (par exemple, la médecine) ou à un objet (par exemple, l'environnement) en particulier. L'éthique est donc générale, et l'éthique appliquée est particulière.

L'éthique appliquée à l'IA serait ainsi une manière d'adapter les doctrines éthiques de la vertu (arétique), conséquentialiste ou déontologique aux problèmes suscités par le développement de l'IA[2]. Deux types d'enjeux peuvent alors être traités : l'enjeu du bon comportement des personnes qui ont un rapport à l'IA, que ce soit dans les phases de recherche, de développement ou de déploiement, et l'enjeu du bon comportement des machines, qu'on leur prête une agentivité morale ou non. Un autre type d'enjeu doit toutefois être ajouté à ces deux-là, un enjeu qui se trouve à la frontière du problème de la gouvernance de l'IA : celui du bon comportement des institutions.

Pour faire de l'éthique appliquée, pour chacun de ces enjeux, il faudrait appliquer des principes préalablement élaborés dans le cadre des grandes doctrines morales. Ainsi, dans le cas du dilemme du tramway, si on privilégie l'éthique conséquentialiste, on dira que la machine se comporte de manière éthique si elle choisit de tuer le moins de personnes ; que que l'informaticienne ou l'informaticien s'est comporté de manière éthique s'il a programmé ou entraîné la machine pour qu'elle tue le moins de personnes ; et que l'institution (par exemple, l'État) est bonne (a des préoccupations éthiques) si elle a encadré le développement et le déploiement des véhicules autonomes de sorte que seules les machines conséquentialistes se retrouvent sur les routes.

2 Voir «The Ethics of the Ethics of AI » (Powers & Ganascia, 2020) à propos des défis posés par l'éthique appliquée à l'IA.

### 2.1.2 L'origine de l'éthique appliquée : le cas de la bioéthique

Cela dit, cette illustration d'une éthique appliquée n'est vraie que si l'on fait l'hypothèse que les normes éthiques seront appliquées sans modification. Or on peut penser qu'elles se modifieront au contact de l'objet de la réflexion. Par exemple, l'éthique appliquée aux patients n'est pas la même chose que l'éthique normative générale : parler de respect de la dignité des patients prend son sens dans le contexte particulier d'une relation avec des professionnels de santé.

Prenons un cas simple : faut-il dire la vérité à un patient sur son état de santé s'il n'a aucune chance de vivre plus de deux mois ? Une première possibilité pourrait être d'appliquer le principe conséquentialiste sur la base que mentir augmentera le bien-être (Collins, 1927). Cette option présente toutefois des difficultés de calcul de conséquences incertaines. Une autre possibilité pourrait être d'appliquer le principe déontologique de respect de l'autonomie des personnes. Cependant, le sentiment que ce choix pourrait faire inutilement du mal persisterait malgré tout. Dans une telle situation, l'éthicienne ou l'éthicien doit prendre en considération des circonstances banales qui deviennent pourtant cruciales pour trouver une solution éthique : il faut dire la vérité, mais il faut la dire d'une manière qui soit bienveillante et attentive à la vulnérabilité du patient (Higgs, 2007).

## L'ÉTHIQUE MÉDICALE ET LA BIOÉTHIQUE

Historiquement, l'éthique médicale et la bioéthique ont été les premières formes d'éthique appliquée. Elles sont nées du traumatisme des expérimentations nazies sur les déportés dans les camps de concentration[3]. Au lendemain de la Seconde Guerre mondiale, le procès des médecins de Nuremberg, terminé en 1947, a donné lieu à un jugement comprenant le Code de Nuremberg, qui établit 10 principes de l'expérimentation éthique, dont celui de consentement éclairé. L'Association Médicale Mondiale s'appuiera sur le Code de Nuremberg en 1964 quand elle adoptera la Déclaration d'Helsinki, dans laquelle apparaît le principe de bienfaisance. L'éthique médicale et la bioéthique se constituent dès lors en une branche à part entière de l'éthique normative reposant sur quatre principes fondamentaux : la bienfaisance, la non-malfaisance, l'autonomie et la justice.

3 Les préoccupations éthiques en médecine sont évidemment plus anciennes; on les trouve déjà chez Hippocrate.

## 2.2. L'éthique de l'IA contre « l'algorithméthique »

Deux leçons sur l'éthique de l'IA peuvent être tirées de ce rappel historique. Premièrement, l'éthique médicale se présente comme une synthèse des différentes doctrines morales. En l'absence de consensus sur une valeur suprême, il faut mettre en équilibre les principes de la vertu (arétique), conséquentialiste et déontologique. La bioéthique fournit alors un modèle pour l'éthique appliquée en général et pour l'éthique de l'IA en particulier. Ainsi, la Déclaration de Montréal pour un développement responsable de l'IA (2018), l'un des tout premiers documents en éthique de l'IA, s'inscrit dans le sillage de la Déclaration d'Helsinki. Dans la Déclaration de Montréal, des principes qui n'avaient encore jamais été utilisés dans le domaine de l'éthique sont proposés pour l'IA, comme le principe de solidarité humain-machine et celui de développement durable, qui prend en considération la matérialité de l'IA et ses impacts sur les conditions naturelles et sociales d'existence des humains, des animaux et du vivant en général. Ce principe de développement durable ne semblait pas un enjeu en bioéthique ; il devient central en éthique de l'IA.

Deuxièmement, il n'existe pas de prêt-à-penser en éthique, il n'existe ni formule ni algorithme qui permettent d'en appliquer les principes. Certains voudraient « algorithmiser l'éthique » (faire une « algorithméthique ») pour que les algorithmes puissent raisonner éthiquement. Or l'analyse d'une situation morale (comme celle du dilemme du tramway ou du véhicule autonome) s'appuie sur un effort de réflexion. Cet effort de réflexion permet de formuler des principes éthiques, de définir les caractéristiques essentielles de la situation, puis d'appliquer les principes pertinents selon une méthode de pondération ou de hiérarchisation. **Ainsi, face aux interrogations légitimes des informaticiennes et informaticiens sur la manière d'appliquer les principes, une seule réponse est possible : il faut réfléchir en situation et se garder de vouloir automatiser le raisonnement moral.** C'est exactement dans ce cadre que l'interdisciplinarité (ou la transdisciplinarité) est essentielle : elle permet d'enrichir la réflexion morale des des développeuses et des développeurs et celle des responsables qui prennent la décision de déployer des systèmes d'intelligence artificielle qui ont des impacts sociaux.

« ...il n'existe pas de prêt-à-penser en éthique, il n'existe ni formule ni algorithme qui permettent d'en appliquer les principes. »

# 3
# Les méthodes de l'éthique de l'IA

*Le rôle de l'éthicien dans une démocratie n'est pas de régler nos plus graves problèmes d'éthique, mais de les éclairer pour que le débat démocratique puisse se faire dans des termes adéquats qui cernent véritablement le (ou les) nœud(s) du problème* (Weinstock, 2006).

# 3. Les méthodes de l'éthique de l'IA

L'éthique est une recherche du sens moral qui oriente l'existence et organise la vie sociale. La réflexion éthique se présente alors comme un effort pour nous comprendre nous-mêmes, nous connaître en tant que collectivité, et pour dégager les valeurs qui sous-tendent notre vie commune. Il ne faut pas penser que ces valeurs sont particulières à une culture au point qu'elles seraient incompréhensibles aux autres communautés. Les valeurs sont plutôt ces points de référence que toute communauté partage, mais qui peuvent être appréciés, organisés et hiérarchisés d'une manière propre à chacune. Les questions que nous nous posons comme membres d'une communauté morale portent sur notre identité comme communauté (qui sommes-nous ?), sur le genre de société que nous voulons pour toutes et tous, sur le type de solidarité que nous voulons construire. Ce questionnement est fondamental et constitue le point de départ de toute réflexion pour bien ancrer la manière de travailler l'éthique.

## 3.1. L'éthique délibérative : l'exemple de la Déclaration de Montréal

Pour déterminer dans quelle société nous voulons vivre et sur quels principes éthiques nous voulons que l'IA se base, il est indispensable de recourir à une forme de réflexion collective, raisonnable et inclusive qui est la méthode de la délibération. La délibération est une manière de parvenir à une décision (un jugement pratique) par un échange d'arguments rationnels et raisonnables. L'échange d'arguments se déroule dans une conversation entre des individus qui vivent dans un même contexte social, mais n'ont pas le même vécu, le même point de vue, les mêmes conceptions de la vie et du bien. Néanmoins, dans cette conversation, chacun essaie de comprendre les autres et de produire la meilleure compréhension du problème discuté. Dans ce processus délibératif, les participants s'efforcent de définir les valeurs qui sous-tendent leur vie sociale. La délibération exige la pluralité des points de vue pour ouvrir le champ des possibles afin de construire les communs qui permettront de mieux définir ce nouveau vivre-ensemble avec la technologie. C'est en quelque sorte une intervention sur la transformation tant potentielle que réelle du monde social qui fait appel à une diversité d'acteurs : les parties prenantes multiples (*multistakeholder*) et même toutes les parties (*omnistakeholder*), incluant les citoyennes et citoyens et y compris celles et ceux qui n'utilisent pas de l'IA.

« La délibération exige la pluralité des points de vue pour ouvrir le champ des possibles afin de construire les communs qui permettront de mieux définir ce nouveau vivre-ensemble avec la technologie. »

Cette méthode a été employée pour élaborer la Déclaration de Montréal pour un développement responsable de l'IA (2018). Le processus délibératif inclusif alors utilisé a créé un précédent dans la manière de travailler l'éthique de l'IA en développant la capacité des citoyens de réfléchir aux enjeux soulevés par l'IA. Comme nous le rappelle Daniel Weinstock : **« Le rôle de l'éthicien dans une démocratie n'est pas de régler nos plus graves problèmes d'éthique, mais de les éclairer pour que le débat démocratique puisse se faire dans des termes adéquats qui cernent véritablement le (ou les) nœud(s) du problème » (2006).** Le processus délibératif ne clôt pas la recherche des principes et des normes éthiques, mais il lui donne une direction, un sens. Les éthiciennes et éthiciens prolongent ensuite la délibération avec leur expertise en s'efforçant de formaliser et de rendre plus cohérentes les propositions issues des délibérations. Le processus démocratique dans lequel s'inscrit la délibération a permis de débattre des principes éthiques et de démontrer la pertinence de cette approche : « C'est à l'intelligence humaine et collective de définir les finalités de la vie sociale et en fonction d'elles, les orientations du développement de l'intelligence artificielle afin qu'il soit socialement responsable et moralement acceptable» (Déclaration de Montréal, 2018). Cette démocratisation, ou modification, du lieu de pouvoir quant à la prise de décision technologique prend davantage en compte l'action, les besoins et les valeurs des sociétés (Feenberg, 1999). L'essentiel est de créer des conditions propices pour la mise en place d'une délibération ouverte et engagée sur les questions technologiques.

Il est clair que le développement des capacités de délibération, d'innovation et d'adaptation technologiques est vital pour le progrès social, mais il doit être couplé à des mécanismes participatifs, impliquant les citoyennes et citoyens, qui favorisent un processus dynamique d'apprentissage de la technologie. Cela implique la création d'espaces sociaux consultatifs où les communautés peuvent évaluer les besoins et les options technologiques ainsi que les impacts de celles-ci sur leur communauté. Les conditions favorisant la délibération sont importantes pour donner aux individus et aux communautés les moyens de faire des choix significatifs en matière de technologie, de passer du statut d'utilisateurs ou de sujets technologiques passifs à celui d'agents actifs qui façonnent de manière constructive les modèles de développement technologique. Une telle démarche offre aux citoyennes et citoyens des moyens d'émancipation, de formation et de capacitation plutôt que d'en faire les cobayes d'expériences technologiques (Latour, 2001).

« Cela implique la création d'espaces sociaux consultatifs où les communautés peuvent évaluer les besoins et les options technologiques ainsi que les impacts de celles-ci sur leur communauté. »

# Conclusion

L'IA comme ensemble de techniques numériques transforme en profondeur certaines pratiques sociales et nous oblige à reprendre les questions éthiques à l'aune des valeurs sociétales et des priorités que nous voulons mettre de l'avant. L'éthique de l'IA offre une occasion unique de penser ensemble le particulier, le singulier, le général, l'irréductible, tout en construisant ce vivre-ensemble. L'époque actuelle est critique. Elle nous offre une occasion unique de travailler à l'établissement d'une plus grande justice sociale et de le faire en exploitant au maximum les possibilités d'une délibération collective qui tient compte des avancées scientifiques et des besoins réels des sociétés. C'est la contribution que peut apporter l'éthique à cette exigence d'harmonisation sociétale afin que toutes et tous puissent vivre décemment et profiter des nouvelles technologies.

« L'éthique de l'IA offre une occasion unique de penser ensemble le particulier, le singulier, le général, l'irréductible, tout en construisant ce vivre-ensemble. »

# Pistes de recommandation

1

**Instaurer des formations en éthique** de l'IA pour le secteur des sciences et technologies en incluant la perspective des sciences sociales et humaines.

2

**Favoriser le développement de la « compétence éthique »** des chercheuses et chercheurs et développeuses et développeurs en IA, de même que celle des législatrices et législateurs ainsi que des autres actrices et acteurs impliqués dans le développement, la commercialisation, l'utilisation et la régulation de l'IA.

3

**Sensibiliser les citoyennes et citoyens** aux différents impacts sociétaux et enjeux éthiques liés au développement et déploiement de l'IA via des conférences, ateliers et autres activités grand public.

4

**Mettre en place des lieux, instances et autres mécanismes** favorisant la réflexion, la discussion et l'évaluation des technologies issues de l'IA et leur utilisation et ce, dans une perspective interdisciplinaire et participative impliquant les citoyennes et citoyens ainsi que les organisations de la société civile.

5

**Instaurer des comités d'éthique interdisciplinaires** spécifiques à l'IA et aux technologies numériques et robotiques au niveau universitaire ainsi qu'au sein des organismes publics.

# Lexique

**Intelligence artificielle générative :** «technologie d'intelligence artificielle (IA) qui génère automatiquement du contenu en réponse à des demandes (*prompts*) rédigés dans des interfaces conversationnelles en langage naturel. Au lieu de simplement effectuer une curation des pages web existantes, en s'appuyant sur le contenu existant, l'IA générative produit plutôt du nouveau contenu. Le contenu peut apparaître dans des formats qui comprennent toutes les représentations symboliques de la pensée humaine : textes écrits en langage naturel, images (y compris des photographies, des peintures numériques et des dessins animés), vidéos, musique et code logiciel » [Notre traduction]( (UNESCO, 2023).

**Système d'intelligence artificielle (SIA):** « système automatisé qui, pour un ensemble donné d'objectifs définis par l'homme, est en mesure d'établir des prévisions, de formuler des recommandations, ou de prendre des décisions influant sur des environnements réels ou virtuels. Pour ce faire, il se fonde sur des entrées machine et/ou humaines pour percevoir les environnements réels et/ou virtuels ; transcrire ces perceptions en modèles (par des moyens automatisés, en s'appuyant par exemple sur l'apprentissage automatique, ou manuels) ; et utiliser des inductions des modèles pour formuler des possibilités de résultats (informations ou actions à entreprendre). Les systèmes d'IA sont conçus pour fonctionner à des niveaux d'autonomie divers » (OCDE, 2019).

**Long-termisme:** Idée soutenant que l'influence positive du futur à long terme, voire à très long terme, est une priorité morale majeure de notre époque et qu'il faut prioriser la prévention des risques existentiels pour l'humanité, notamment dans l'allocation des ressources (Boddington, 2023; MacAskill, 2022). En s'appuyant sur le postulat que la très vaste majorité des êtres humains ne sont probablement pas encore nés (MacAskill, 2022), et en se basant ensuite sur un raisonnement conséquentialiste, les long-termistes prétendent que la survie de la civilisation à très long terme, en permettant l'existence de ces plusieurs milliards de futures vies, décuplerait ultimement la valeur totale produite dans le monde (Boddington, 2023).

**Transhumanisme:** «Mouvement intellectuel et culturel soutenant la possibilité et la désirabilité d›améliorer fondamentalement la condition humaine par la raison appliquée, notamment en développant et en diffusant largement les technologies permettant d›éliminer le vieillissement et d›améliorer considérablement les capacités intellectuelles, physiques et psychologiques de l›être humain. » [Notre traduction](Bostrom, 2003).

**Singularitarisme:** Mouvement apparenté au transhumanisme postulant l'avènement de la singularité (Cassiani-Laurin, 2018), c'est-à-dire un moment de l'histoire où l'évolution accélérée de la technologie, et notamment de l'IA, aurait des impacts si profonds qu'elle transformerait de manière permanente la vie humaine, principalement par la fusion de l'humain avec la machine (Kurzweil 2005).

# Bibliographie


Awad, E., Dsouza, S., Kim, R. & al. (2018). The Moral Machine experiment. *Nature,* 563(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Boddington, P. (2023). Towards the Future with AI: Work and Superintelligence. Dans *AI Ethics. Artificial Intelligence: Foundations, Theory, and Algorithms*. Springer, Singapore. https://doi-org.acces.bibl.ulaval.ca/10.1007/978-981-19-9382-4_10

Bostrom, N. (2003). The Transhumanist FAQ: A General Introduction. (Version 2.1). *World Transhumanist Association.* https://nickbostrom.com/views/transhumanist.pdf

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S.M., Frith, C., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M.A.K., Schwitzgebel, E., Simon, J. & VanRullen, R. (2023). Consciousness in Artificial Intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*. https://doi.org/10.48550/arXiv.2308.08708

Cassiani-Laurin, F. (2018). *La singularité transhumaniste comme mythe eschatologique contemporain : l'utopie technologique en tant qu'achèvement du projet moderne.* [mémoire de maîtrise, Université du Québec à Montréal]. Archipel. https://archipel.uqam.ca/11714/

Center for AI Safety. (2023). *Statement on AI Risk*. https://www.safe.ai/statement-on-ai-risk

Collins, J. (1927). Should doctors tell the truth?. *Harper's Monthly Magazine*, 155, 320-6.

Déclaration de Montréal. (2018). La Déclaration de Montréal pour un développement responsable de l'intelligence artificielle. https://declarationmontreal-iaresponsable.com/la-declaration/

Dratwa, J. (2019). *Dans quel monde voulons-nous vivre ensemble?: éthique et Europe* (Vol. 4). ISTE.

European Group on Ethics in Science and New Technologies (EGE). (2018). Statement on Artificial Intelligence, Robotics and "Autonomous" Systems. https://www.unapcict.org/resources/ictd-infobank/statement-artificial-intelligence-robotics-and-autonomous-systems

Feenberg, A. (1999). *Questioning Technology.* Routledge

Floridi, L. (2023). *L'éthique de l'intelligence artificielle : Principes, défis et opportunités.* Mimesis

Floridi, L. & Clement-Jones, T. (2019). The five principles key to any ethical framework for AI. *Tech New Statesman.*

Foot, P. (2002). The Problem of Abortion and the Doctrine of the Double Effect. Dans *Virtues and Vices: and other essays in moral philosophy*, Oxford University Press. https://doi-org.acces.bibl.ulaval.ca/10.1093/0199252866.003.0002. (œuvre originale publiée en 1967).

Georgieva, I., Lazo, C., Timan, T., & van Veenstra, A. F. (2022). From ai ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience. *AI and Ethics*, 2(4), 697–711. https://doi.org/10.1007/s43681-021-00127-3

Higgs, R. (2007). Truth Telling. Dans R. Rhodes, L.P. Francis et A. Silvers (dir.), *The Blackwell Guide to Medical Ethics* (p.88-103). Blackwell Pub. DOI:10.1002/9780470690932

Koniakou, V. (2023). From the "rush to ethics" to the "race for governance" in Artificial Intelligence. *Information Systems Frontiers*, 25(1), 71-102.

Kurzweil, R. (2005). *The singularity is near : when humans transcend biology*. Viking.

Latour, B. (2001). *L'espoir de pandore : pour une version réaliste de l'activité scientifique*. La Découverte et Syros.

MacAskill, W. (2022). What is longtermism?. *BBC*. https://www.bbc.com/future/article/20220805-what-is-longtermism-and-why-does-it-matter

Maclure, J. (2020). The new AI spring: a deflationary view. *AI & Society : Journal of Knowledge, Culture and Communication*, 35(3), 747–750. https://doi.org/10.1007/s00146-019-00912-z

Ménissier, T. (2023). Les quatre éthiques de l'intelligence artificielle. *Revue d'anthropologie des connaissances* [En ligne], 17(2). https://doi.org/10.4000/rac.29961

Nussbaum, M. C. (2012). *Philosophical interventions : reviews, 1986-2011*. Oxford University Press.

OCDE. (2019). *L'intelligence artificielle dans la société*, Éditions OCDE. https://doi.org/10.1787/b7f8cd16-fr

Peres, R., Schreier, M., Schweidel, D., & Sorescu, A. (2023). On ChatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice. *International Journal of Research in Marketing*. https://doi.org/10.1016/j.ijresmar.2023.03.001

Powers, T.M. & Ganascia, JG. (2020). The Ethics of the Ethics of AI. Dans Dubber, M.D., Pasquale, F. & Das, S. (dir.), *The Oxford Handbook of Ethics of AI* (Oxford Academic, p.26-51), Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.2

Ricoeur, P. (2004). *Parcours de la reconnaissance : trois études* (Ser. Les essais). Stock.

UNESCO. (2021). Recommandation sur l'éthique de l'intelligence artificielle. https://unesdoc.unesco.org/ark:/48223/pf0000381137_fre

UNESCO. (2023). Guidance for generative AI in education and research. https://unesdoc.unesco.org/ark:/48223/pf0000386693

Verhaegen, J. (1984). Notions floues et droit pénal. Dans C. Perelman & al.(dir.), *Les notions à contenu variable en droit* (p. 7-19). Bruylant.

Weinstock, D. (2006). *Profession, éthicien*. Presses de l'Université de Montréal.

obvia

www.obvia.ca